3936–3945 *Nucleic Acids Research, 2016, Vol. 44, No. 8* *Published online 2 March 2016*
doi: 10.1093/nar/gkw133

Structures of an all-α protein running along the DNA major groove

Li-Yan Yu[1,†], Wang Cheng[1,†], Kang Zhou[1,†], Wei-Fang Li[1], Hong-Mei Yu[1], Xinlei Gao[1], Xudong Shen[2], Qingfa Wu[1], Yuxing Chen[1,*] and Cong-Zhao Zhou[1,*]

[1]Hefei National Laboratory for Physical Sciences at the Microscale and School of Life Sciences, University of Science and Technology of China, Hefei, Anhui 230027, China and [2]School of Information Science and Technology, University of Science and Technology of China, Hefei, Anhui 230027, China

Received November 11, 2015; Revised February 22, 2016; Accepted February 23, 2016

ABSTRACT

Despite over 3300 protein–DNA complex structures have been reported in the past decades, there remain some unknown recognition patterns between protein and target DNA. The silkgland-specific transcription factor FMBP-1 from the silkworm *Bombyx mori* contains a unique DNA-binding domain of four tandem STPRs, namely the score and three amino acid peptide repeats. Here we report three structures of this STPR domain (termed *Bm*STPR) in complex with DNA of various lengths. In the presence of target DNA, *Bm*STPR adopts a zig-zag structure of three or four tandem α-helices that run along the major groove of DNA. Structural analyses combined with binding assays indicate *Bm*STPR prefers the AT-rich sequences, with each α-helix covering a DNA sequence of 4 bp. The successive AT-rich DNAs adopt a wider major groove, which is in complementary in shape and size to the tandem α-helices of *Bm*STPR. Substitutions of DNA sequences and affinity comparison further prove that *Bm*STPR recognizes the major groove mainly via shape readout. Multiple-sequence alignment suggests this unique DNA-binding pattern should be highly conserved for the STPR domain containing proteins which are widespread in animals. Together, our findings provide structural insights into the specific interactions between a novel DNA-binding protein and a unique deformed B-DNA.

INTRODUCTION

Recognitions of proteins towards specific DNA sequences are indispensable to read out the genetic information for all living organisms. Since the first X-ray structure of protein–DNA complex reported in 1987 (1), we have illustrated more and more structural insights into how a protein selectively binds to one or a few DNA sites out of millions along the genome. The previous proposal of 'simple recognition code' has been proved to be inaccurate to describe the specific interactions between protein and DNA (2–4). Instead, structural analyses reveal that specific recognitions of protein towards DNA are accomplished by the combination of both base (direct) readout and shape (indirect) readout (5–7). The former is involved in direct interactions, such as hydrogen bonds and/or hydrophobic contacts between amino acids and nucleotide bases (7–9), whereas the latter corresponds to the recognition of protein towards sequence-dependent DNA conformation, such as the curvature and narrow minor groove of A-tracts (10–12). To date, >3300 DNA-complexed protein structures are available in the database (http://npidb.belozersky.msu.ru/) (13,14), which are grouped into ∼100 superfamilies according to Structural Classification of Proteins (15). However, most protein–DNA interaction patterns are dominantly mediated by base readout, whereas the cases mainly or exclusively contributed by DNA shape readout are relatively rare.

The silkglands of silkworm *Bombyx mori* have been known as the most efficient factories in nature that produce the silk proteins (16). In the posterior silkglands, the fibroin gene is selectively transcribed at the fifth instar larval stage (17). A series of transcriptional factors, which were originally identified from the crude extract of posterior silkglands, finely coordinate this efficient expression system via specifically binding to the regulatory elements at the upstream and/or the intron of fibroin gene (18–21). Remarkably, the fibroin modulator binding protein-1 (FMBP-1) possesses three binding elements around −130, +220 and +290 sites of fibroin gene (21). It exhibits a tissue- and stage-specific expression profile that perfectly correlates with that of fibroin gene (21–23). Sequence analysis reveals the 218-residue FMBP-1 consists of two distinct domains (Figure 1A). The N-terminal domain of unknown function con-

[*]To whom correspondence should be addressed. Tel: +86 551 63602492; Fax: +86 551 63600406; Email: zcz@ustc.edu.cn
Correspondence may also be addressed to Yuxing Chen. Tel: +86 551 63602492; Fax: +86 551 63600406; Email: cyxing@ustc.edu.cn
[†]These authors contributed equally to the paper as first authors.

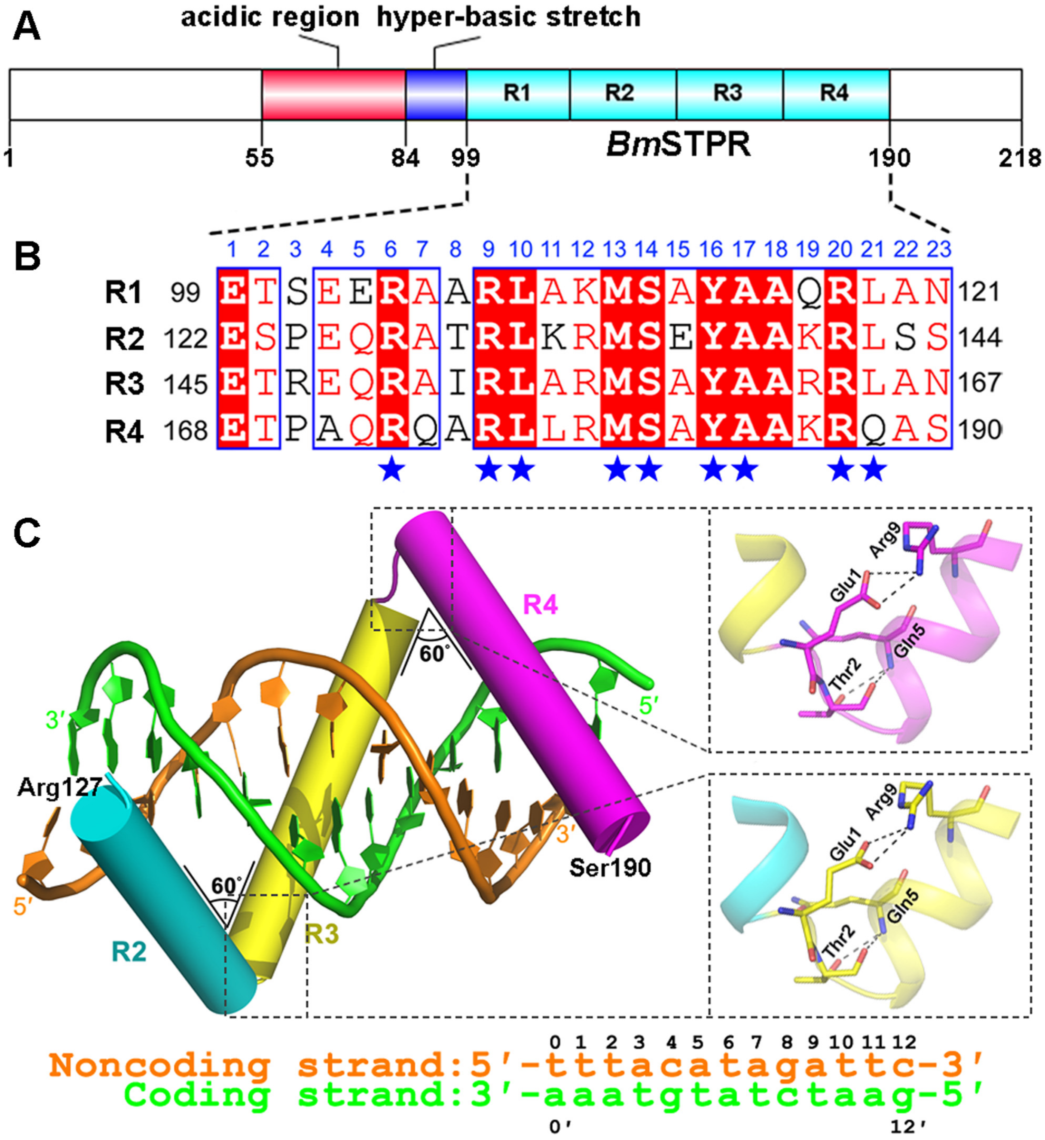

Figure 1. Structure of *Bm*STPR. (**A**) Domain organization of FMBP-1. (**B**) Sequence alignment of the four repeats of *Bm*STPR. The highly conserved residues involved in DNA binding are labelled with blue pentangles. (**C**) Crystal structure of *Bm*STPR in complex with the 13-bp DNA. Repeats R2 to R4 are shown as cylinders and coloured in cyan, yellow and purple, respectively. The coding and noncoding strands of the 13-bp DNA are shown as green and orange, respectively. The detailed interactions that stabilize the repeats R3 and R4 of *Bm*STPR are zoomed-in at the right panel. The involved residues are labelled and shown as sticks.

tains an acidic region (residues Glu55–Glu84) followed by a hyper-basic stretch (residues Pro85–Ser98), whereas the C-terminal DNA-binding domain (residues Glu99–Thr218) consists of four tandem repeats R1–R4, each of which contains 23 residues, thus termed the score and three amino acid peptide repeat (STPR) (24). The four repeats of this domain (termed *Bm*STPR for short) are highly homologous to each other with a sequence-identity of 60–80% (Figure 1B), which was proposed to favour DNA fragments with a consensus sequence of 5′-atntwtnta-3′ (n: any nucleotide, w: a or t) through cooperative binding (24). In the absence of DNA, only the N-terminal moiety of each repeat of *Bm*STPR is folded into a short α-helix (25), whereas the intact repeat adopts an α-helical structure upon the addition of a hydrogen-bond promoting solvent trifluoroethanol (25,26). Competitive binding assays further suggested that *Bm*STPR most likely binds to the major groove of DNA (27). Bioinformatic analysis indicated the STPR domain is widespread in diverse eukaryotic organisms, including the model organisms *Caenorhabditis elegans*, *Drosophila*, mouse and human (24). However, the DNA-binding profile of the STPR domain remains unknown due to the lack of its intact structure in complex with DNA.

Here we present three structures of *Bm*STPR in complex with DNA of various lengths. Upon binding to either a 13-bp or 20-bp DNA fragment derived from the +290 site of fibroin gene, repeats R2–R4 of *Bm*STPR fold into three tandem α-helices, running along the major groove of

DNA, whereas all or the majority of R1 is missing in the electron density map. The three repeats display a relatively rigid helical structure, forming an inter-helix angle of about 60°, exactly covering a 4-bp DNA segment by each repeat. This regular binding pattern enabled us to design a double-stranded DNA containing four tandem repetitive units of 5′-atac-3′, which makes the intact R1 fold into a helix similar to that of R2–R4. Biochemical study indicated that *Bm*STPR favours the AT-rich sequences, which most likely adopts a narrower minor groove (28,29), and a wider major groove to accommodate the rigid α-helix of *Bm*STPR. Notably, the DNA bound to *Bm*STPR adopts a unique deformed B-DNA conformation. Moreover, substitutions of DNA sequences combined with binding assays reveal that *Bm*STPR recognizes the DNA major groove mainly via indirect interactions. Together, our findings provide structural insights into a novel protein–DNA interaction pattern that mainly mediated by DNA shape readout.

MATERIALS AND METHODS

Samples preparation

The coding region of *Bm*STPR (residues Glu99–Ser193) was cloned into the ligation-independent cloning vector 2BT with an N-terminal 6×His tag. The construct was overexpressed in *Escherichia coli* BL21 (DE3) strain (Novagen) at 37°C for 4 h after induction by 0.2 mM isopropyl β-D-l-thiogalactopyranoside at an OD_{600nm} of 0.8. Cells were harvested and resuspended in the lysis buffer (1 M NaCl, 20 mM potassium phosphate, pH 9.0), and then disrupted by sonication. After centrifugation, the His-tagged fusion proteins were isolated with Ni-NTA affinity column (Qiagen) and further purified by gel filtration (Superdex 75, GE Healthcare) in a buffer containing 1 M NaCl, 20 mM Tris-HCl, pH 9.0. The peak fractions containing the target protein were collected and then applied to the desalting column (Hiprep 26/10, GE Healthcare) in the buffer containing 7.5 mM $MgCl_2$, 60 mM NaCl and 30 mM Tris-HCl, pH 7.9. The eluted proteins were pooled and frozen for further study.

The selenium-methionine (SeMet)-labelled *Bm*STPR protein was overexpressed in *E. coli* strain B834 (DE3). Transformed cells were grown at 37°C in SeMet medium (M9 medium supplemented with 25 μg/ml SeMet and other amino acids at 50 μg/ml) to an OD_{600nm} of 0.8, and then induced with 0.2 mM isopropyl β-D-l-thiogalactopyranoside for another 4 h. The *Bm*STPR mutants were obtained with the Mut Express^{TM} Fast Mutagenesis Kit using the plasmid encoding the wild-type *Bm*STPR as the template. SeMet substituted and mutant *Bm*STPR proteins were purified using the same protocol used for the native protein.

Single-stranded DNA (ssDNA) was synthesized by Sangon Biotech (Shanghai). The ssDNA was resuspended in the buffer containing 7.5 mM $MgCl_2$, 60 mM NaCl and 30 mM Tris-HCl, pH 7.9, and then mixed with a complementary strand with equal molar amount. After heating at 95°C for about 6 min, the mixture was annealed by slow cooling to room temperature to prepare the double-stranded DNA.

Crystallization, data collection and processing

The protein–DNA complexes were obtained by incubation of *Bm*STPR with the DNA fragments at a molar ratio of 1:1.2 for 40 min on ice. Afterwards, the mixture was concentrated to ∼18 mg/ml for crystallization at 289 K. The optimized crystals of *Bm*STPR in complex with the 13-bp DNA (5′-tttacatagattc-3′) appeared in the solution containing 20% (v/v) 2-propanol, 17% (w/v) polyethylene glycol 4000 and 0.1 M sodium citrate tribasic dehydrate, pH 5.6. The crystals in complex with the 20-bp DNA (5′-agtatttacatagattcatc-3′) were obtained from the reservoir solution of 16% (v/v) glycerol, 22% (w/v) polyethylene glycol 3350, 0.2 M ammonium citrate tribasic, pH 7.0, whereas the crystals complexed with the 18-bp DNA (5′-catacatacatacataca-3′) were obtained from the solution containing 18% (w/v) polyethylene glycol 2000, 0.1 M sodium citrate tribasic dehydrate, pH 5.6.

The crystals were transferred to a cryoprotectant-containing glycerol and flash-cooled in liquid nitrogen. The diffraction data were collected at 100 K in a liquid nitrogen stream using beamline BL17U with a Q315rCCD (ADSC, MARresearch, Germany) at the Shanghai Synchrotron Radiation Facility. The data were indexed, integrated and scaled with the HKL2000 package (30).

Structure determination and refinement

Using a SeMet-substituted protein crystal, the structure of *Bm*STPR in complex with 13-bp DNA was determined by the single wavelength anomalous dispersion phasing method (31) with the program *phenix.solve* implemented in PHENIX (32). The initial model was built automatically with the program *AutoBuild* in PHENIX. The complete model of *Bm*STPR in complex with 13-bp DNA was built manually using the *Coot* program (33). The model was then refined with the *Refmac5* program (34) and TLS refinement (35). Using the 13-bp DNA complexed structure as the search model, the other two complex structures were determined with molecular replacement and refined with the same procedure. The final models were evaluated with the programs *MolProbity* (36) and *Procheck* (37). Data collection and structure refinement statistics are listed in Supplementary Table S1. All structure figures were prepared using the program *PyMOL* (38).

Logo formation of the repetitive units favoured for *Bm*STPR

Sequence logos were generated with the *seqLogo* software (39), which is used for graphical representation of nucleic acids for displaying the patterns in a set of aligned sequences. We first used the context-independent algorithm, where the probability of the 4-bp repetitive unit x(1)x(2)x(3)x(4) is calculated by the formula $P_{x(1)x(2)x(3)x(4)} = P_{x(1)} \times P_{x(2)} \times P_{x(3)} \times P_{x(4)}$ [x(n) is the nucleotide at the position n of the 4-bp unit]. The weight of each repetitive 4-bp unit of 135 possible combinations was given with the value equal to the relative folds of its binding affinity to that of 5′-$(gcca)_3$-3′, which has the lowest binding affinity towards *Bm*STPR. The correlation analysis revealed a value of 0.658 with a *P*-value <2.2e-16. The related seqLogo graphic was shown as Supplementary Figure S3. Alternatively, a context-dependent model was generated with the first-order Markov chain algorithm, where the probability of each 4-bp unit is calculated with the formula $P_{x(n+1)} = P_{x(n)} \times P_{\text{transition matrix}}$ [$P_{\text{transition matrix}}$ is the probability of the transition from base x(n) to x(n+1)]. The transition probability matrix was generated according to the *K*d values of the 135 DNA sequences of three 4-bp repetitive units. The correlation analysis shows a much higher value of 0.824 with a *P*-value <2.2e-16. The related seqLogo graphic was shown as Figure 3.

Isothermal titration calorimetry (ITC)

Microcalorimetric titrations were performed at 25°C employing a MicroCal iTC_{200} instrument (GE Healthcare). Both samples of protein and DNA were dissolved in the buffer of 7.5 mM $MgCl_2$, 60 mM NaCl and 30 mM Tris-HCl, pH 7.9, and then degassed before use. The sample cell was loaded with 200 μl DNA at 10 μM, whereas the injection syringe was loaded with 40 μl *Bm*STPR at 280 μM. The number and injected volume of the titration steps (0.4 μl+19×2 μl) were the same for all measurements, and the spacing between injections was set to 120 s. Additionally, heats of dilution, determined by titrating the proteins into solution buffer (7.5 mM $MgCl_2$, 60 mM NaCl and 30 mM Tris-HCl, pH 7.9), were subtracted from the raw titration data. Analyses of all data were performed with MicroCal Origin software accompanying the ITC instrument.

RESULTS AND DISCUSSION

Overall structure of *Bm*STPR–DNA

In order to obtain a suitable DNA sequence for co-crystallization with *Bm*STPR, we first compared the binding affinities of three previously reported DNA sequences of 28 bp (21) and found that the fragment derived from +290 site of the fibroin gene displayed the highest affinity towards *Bm*STPR (Supplementary Table S2). Further screening of an optimum DNA length enabled us to focus on two DNA fragments of 13 bp (5′-tttacatagattc-3′) and 20 bp (5′-agtatttacatagattcatc-3′), respectively, which were applied to co-crystallization trials. Eventually, we succeeded in solving the crystal structures of *Bm*STPR complexed with the 13-bp DNA at 1.95 Å and the 20-bp DNA at 2.40 Å.

In the 13-bp DNA complexed structure, only the residues of repeats R2 to R4 (Arg127–Ser190) could be clearly traced in the electron density map (Figure 1C). The three tandem α-helices run along the DNA major groove, in a reverse direction of the fibroin gene coding strand (Figure 1C). Each STPR starts with a two-residue linker followed by a 21-residue helix (residues No. 3–23), with an inter-helix angle of ∼60° (Figure 1C). Similar to the previously reported solution structures of the four individual repeats (25), we also observed salt bridges between the side chains of two highly conserved Glu1 and Arg9 in the repeats R3 and R4, in addition to two hydrogen bonds between the backbone nitrogen of Gln5 and the two oxygen atoms of Thr2 (Figure 1C). These interactions have been proposed to stabilize the α-helical conformation of the N-terminal moiety of each repeat in the absence of DNA (25). In fact, substitution of Glu1 with Gln in any repeat could lead to the decrease of DNA-binding affinity towards *Bm*STPR (25).

Despite in the presence of an extended DNA sequence of 20 bp, the repeat R1 remains partially folded into a short α-helix of six residues (Supplementary Figure S1), indicating that there is no specific interaction between R1 and DNA sequence at +290 site of fibroin gene. As shown in Supplementary Figure S2, the truncated protein without R1 possesses a *K*d value of 546.6 nM, comparable to that of the full-length *Bm*STPR with a *K*d value of 118.8 nM towards the 20-bp DNA. It suggested that the latter three repeats are sufficient for *Bm*STPR to specifically recognize the regulatory element at +290 site of fibroin gene. Notably, we found that R2 and the partially folded R1 also form an inter-helix angle of about 60°, implying that the regular angle between two adjacent helices of *Bm*STPR is an induced fit upon binding to the consecutive DNA major groove.

The tandem binding pattern between *Bm*STPR and DNA

The complex structure of *Bm*STPR with 13-bp DNA that has a much higher resolution was applied to further structural analyses. The repeats R2 to R4 run along the major groove of 12-bp DNA from t1:a1′ to c12:g12′, with each repeat covering 4-bp DNA (Figure 2A). The three base pairs, c4:g4′, g8:c8′ and c12:g12′, facing the sharp turn of two adjacent helices have no interaction with the protein. The highly conserved residues Arg6, Arg9, Leu10, Tyr16 and Arg20 are involved in salt bridges or hydrogen bonds with the phosphate groups, whereas the side chains of the conserved residues Met13, Ser14, Ala17 and Leu21 in R2 and R3 recognize DNA via hydrophobic interactions (Figures 1B and 2). Binding assays also proved that this tandem interaction pattern of 4-bp DNA per α-helix covers the 12-bp DNA from t1 to c12 (Figure 2A), which has a significantly higher affinity compared to the 12-bp DNA from t0 to t11 (Supplementary Table S2).

In detail, Arg6 and Arg9 in the repeats R2 to R4 form salt bridges with the phosphate group of a2′, a6′ and a10′ through their polar side chains, respectively (Figure 2). The main-chain oxygen atom of Leu10 in R3 (or R4) forms a hydrogen bond with the phosphate group of t7′ (or a11′) via a water molecule, whereas Leu10 in R2 shows no interaction with DNA (Figure 2B). Residues Met13, Ser14 and Ala17 of R2 and R3 constitute a hydrophobic pocket to accommodate the methyl group of t3′ and t7′, respectively. However, corresponding hydrophobic contacts are missing between R4 and a11′. The residue Leu21 in R2 (or R3) forms hydrophobic interaction with the methyl group of t5′ (or t9′). The side-chain hydroxyl group of Tyr16 in each repeat of R2 to R4 forms hydrogen bond with the phosphate group of t1, a5 or a9, in addition to hydrophobic contacts with the methyl group of t2, t6 or t10 (Figure 2B). The side chains of Arg20 in R2 and R3 form salt bridges with the phosphate groups of t2 and t6, respectively (Figure 2B). In contrast, the side chain of Arg20 in R4 points towards the DNA major groove and forms a hydrogen bond with the base group of t10 mediated by a water molecule (Figure 2B). To the best of our knowledge, this kind of tandem interaction pattern is unprecedented in previously identified protein–DNA structures.

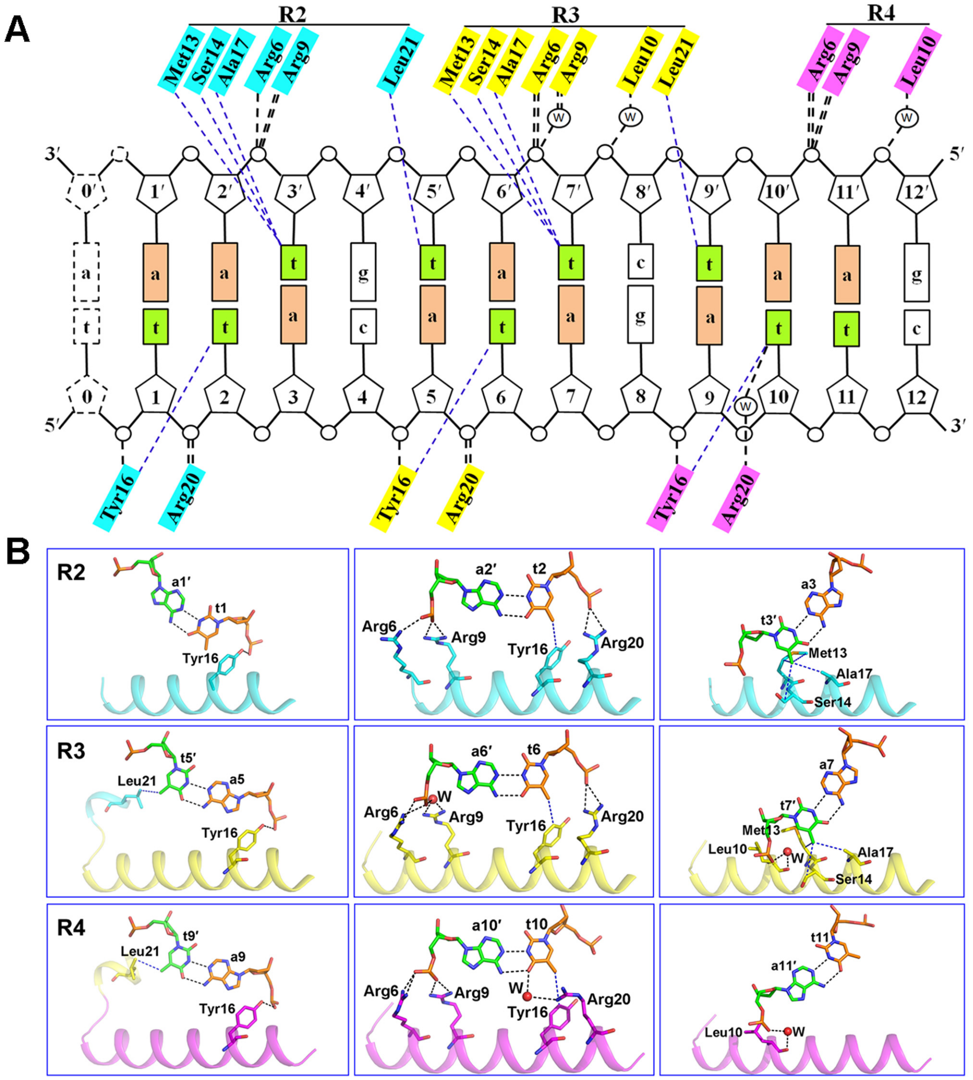

Figure 2. The tandem interactions between *Bm*STPR and 13-bp DNA. (**A**) A diagram of *Bm*STPR interacting with DNA. Residues from R2 to R4 are coloured as their located repeat. Water molecules are donated as open circles labelled with the letter 'W'. The contacted base groups are displayed as light orange and green, respectively. (**B**) Cartoon representation of the contacts between R2 to R4 and corresponding nucleotides. The involved nucleotides and residues are labelled and shown as sticks. Water molecules are shown as red spheres.

The favoured 4-bp DNA repetitive units recognized by *Bm*STPR

The tandem interaction pattern strongly suggested that the highly conserved repeats of *Bm*STPR should be able to bind to the tandem repeats of 4-bp DNA. Accordingly, we synthesized the 12-bp DNA sequences of all 135 possible combinations that contain three 4-bp repetitive units, except for the sequence 5′-(gggg)$_3$-3′ that could not be synthesized, and compared their binding affinity towards *Bm*STPR. Only nine DNA sequences show a lower *K*d value compared to that of the physiologically identified 12-bp DNA (5′-ttacatagattc-3′) (Table 1 and Supplementary Table S3). These sequences are featured with a high A/T content, including six sequences with a 100% A/T repetitive unit (5′-atat-3′, 5′-aata-3′, 5′-attt-3′, 5′-ataa-3′, 5′-tata-3′ or 5′-taaa-3′) and three with 75% A/T (5′-atac-3′, 5′-tatc-3′ or 5′-atag-3′). Notably, the DNA 5′-(gcca)$_3$-3′ possesses a lowest affinity (*K*d of 44563.3 nM), which is about 1% to that of the 12-bp DNA at +290 site.

Based on statistic analyses of these binding affinity data in combination with the first-order Markov chain algorithm, we generated a context-dependent consensus using the *seqLogo* program (39). The consensus is featured with an AT-rich content, with a correlation coefficient value of 0.824 at a *P*-value <2.2e-16 (Figure 3). In contrast, a context-independent consensus also possesses an AT-rich sequence, but exhibits a correlation coefficient value of 0.658 at a *P*-value <2.2e-16 (Supplementary Figure S3). A higher correlation coefficient value of context-dependent logo repeat further indicated that the indirect readout from the context of DNA sequence contributes the majority to binding *Bm*STPR.

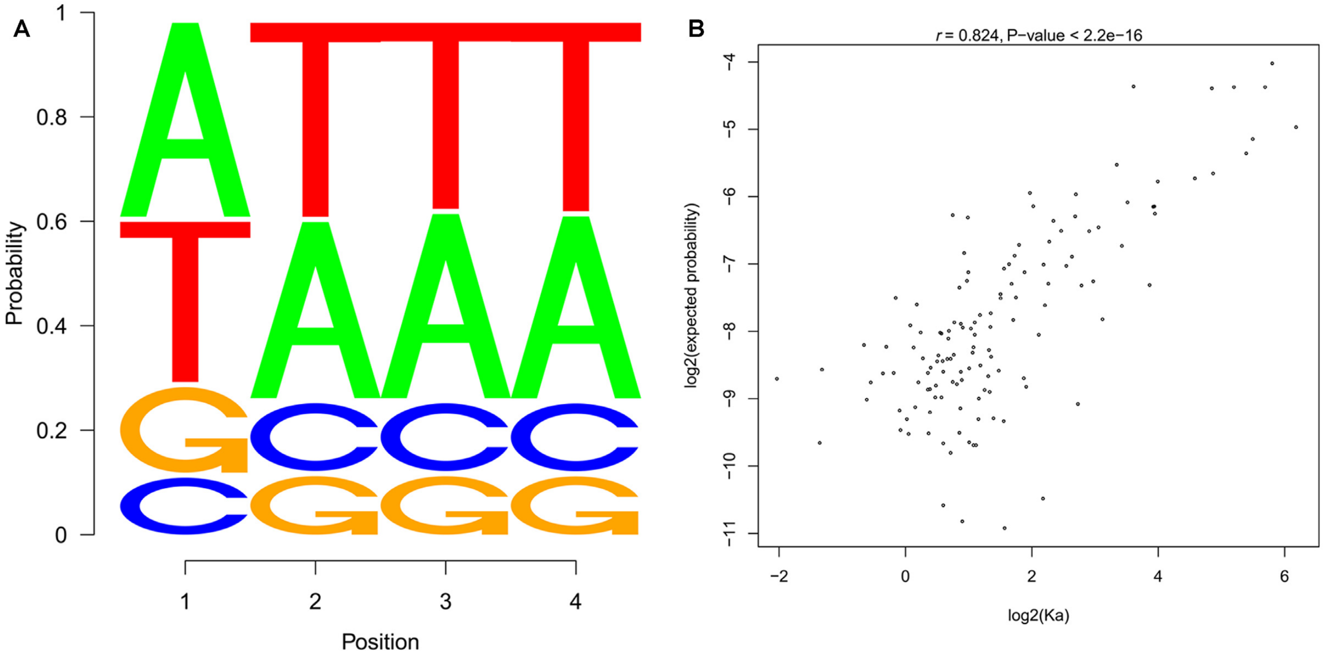

Figure 3. The favoured 4-bp repetitive unit binding to *Bm*STPR. (**A**) A context-dependent consensus generated by the first-order Markov chains algorithm. (**B**) The correlation analysis of the context-dependent consensus.

Table 1. The nine representative DNAs of high binding affinity towards *Bm*STPR

DNA	Sequence (5′→3′)	$n = 3$ *K*d (nM)	$n = 4$ *K*d (nM)
No.1	$(atac)_n$	135.0 ± 7.8	107.5 ± 4.5
No.2	$(atat)_n$	175.1 ± 7.3	97.4 ± 5.7
No.3	$(aata)_n$	189.9 ± 12.6	57.6 ± 3.0
No.4	$(tatc)_n$	216.6 ± 6.1	56.1 ± 3.4
No.5	$(attt)_n$	232.8 ± 8.1	113.8 ± 4.7
No.6	$(ataa)_n$	268.3 ± 25.0	155.7 ± 8.5
No.7	$(atag)_n$	335.4 ± 17.9	199.7 ± 10.3
No.8	$(tata)_n$	340.1 ± 19.6	133.9 ± 8.6
No.9	$(taaa)_n$	412.8 ± 43.8	82.1 ± 8.0
+290	ttacatagattc	422.0 ± 38.8	

Structure of the intact *Bm*STPR in complex with 5′-$(atac)_4$-3′

The 20-bp DNA complexed structure suggested that a target DNA sequence might be able to induce the folding of an intact repeat R1. Using the repetitive units of top nine DNA sequences of highest affinity (Table 1), we synthesized nine sequences of 16-bp DNA composed of four tandem repeats. As expected, binding assays revealed an increased affinity towards *Bm*STPR for all of these 16-bp DNAs (Table 1). Furthermore, we crystalized *Bm*STPR in complex with a 18-bp DNA that contains four repetitive units of 5′-atac-3′ in addition to two protecting nucleotides at both termini, and solved its structure at 2.2 Å. Similar to the 13-bp DNA complexed structure, repeats R2 to R4 of *Bm*STPR wrapping the 18-bp DNA also adopt a 2-residue linker followed by a 21-residue helix (Figure 4A). Moreover, the repeat R1 is indeed folded into a similar helix that lies in the DNA major groove as the other three repeats (Figure 4A). As a result, the four tandem α-helices of *Bm*STPR wrap the 18-bp DNA along the major groove one after another, with an inter-helix angle of 54–63° (Figure 4A). It further suggested that the regular angle between two adjacent helices of *Bm*STPR is resulted from binding to the consecutive DNA major groove.

Structure-based analysis demonstrated that each repeat, including R1, applies an almost identical pattern to wrap a 4-bp DNA, via both direct and indirect contacts (Figure 4B and C). The direct interactions include hydrophobic interactions with the methyl groups of three nucleotide bases t1′, t2 and t3′ of each unit (Figure 4C). For example, t1′ and t2 are separately stabilized by the side chains of Leu21 and Tyr16, whereas the methyl group of t3′ is accommodated in a hydrophobic pocket formed by Met13, Ser14 and Ala17 of each repeat (Figure 4C). To test the contribution of these direct interactions, we substituted the two central thymidylates (namely t2 and t3′ of each 4-bp repetitive unit) with uridylate, respectively. As shown in Supplementary Figure S4, substitution of t2 or t3′ to uridylate in each 4-bp repetitive unit led to a *K*d value of 183.3 or 635.5 nM, which represents a slight decrease of binding affinity as compared to the original 16-bp DNA with a *K*d value of 107.5 nM. In contrast, a single substitution of the central

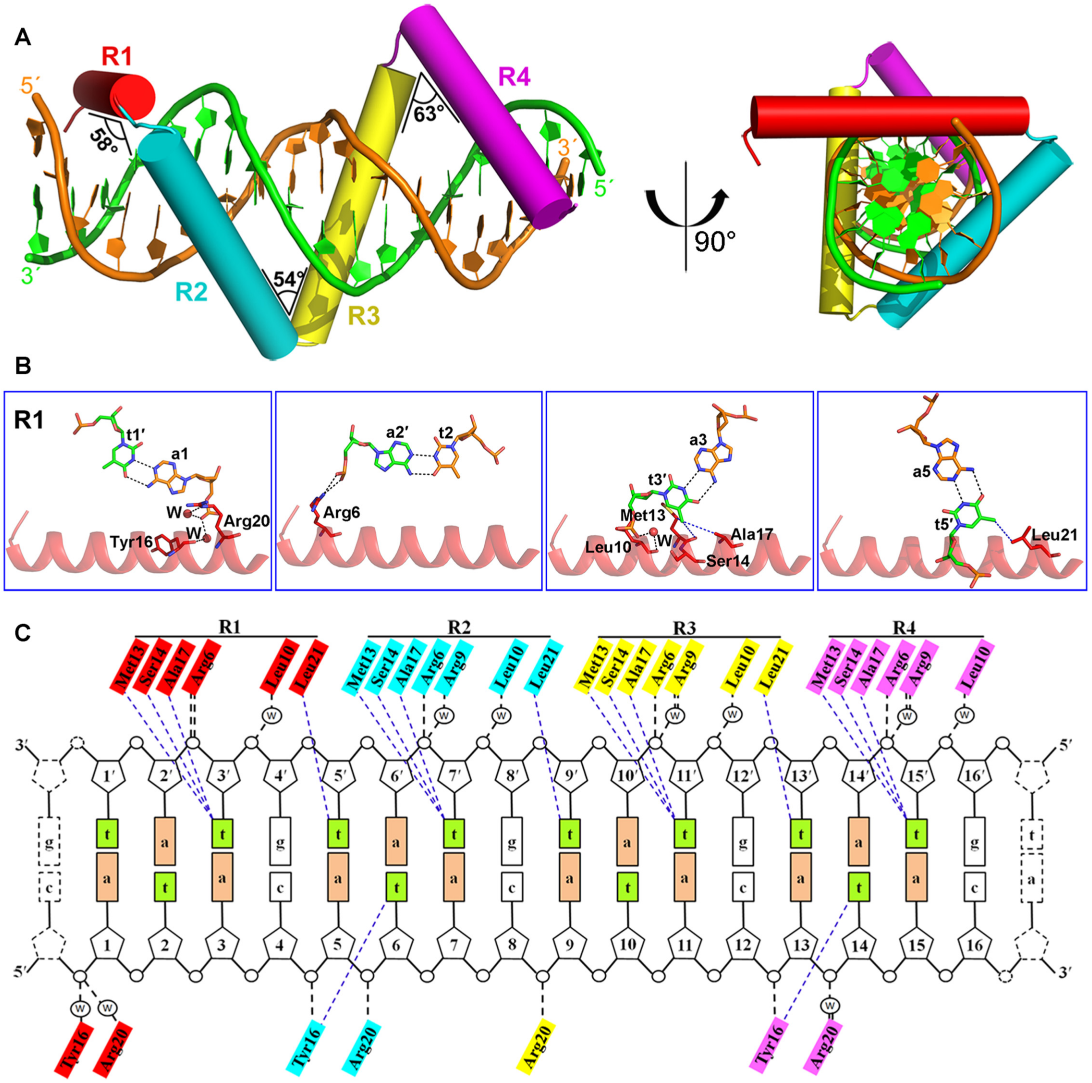

Figure 4. The structure of *Bm*STPR in complex with 18-bp DNA containing four repeats of 5′-atac-3′. (**A**) Cartoon representation of *Bm*STPR in complex with 18-bp DNA. The DNA strands and repeats of *Bm*STPR adopt the same colour coding as Figure 1C, in addition to R1 coloured in red. (**B**) Cartoon representation of the contacts between R1 and corresponding nucleotides in the 18-bp DNA complexed structure. The involved nucleotides and residues are labelled and shown as sticks. The water molecules are indicated as red spheres and marked with the letter 'W'. (**C**) A diagram of the interactions between *Bm*STPR and 18-bp DNA.

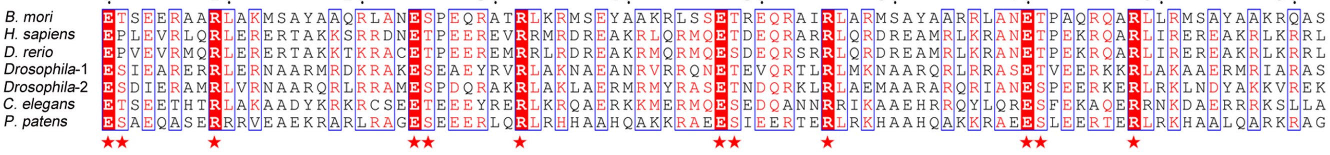
α1
α2
α3
α4
B. mori
1
10
20
30
40
50
60
70
80
90
B. mori ETSEERAARLAKMSAYAAQRLANESPEQRATRLKRMSEYAAKRLSSETREQRAIRLARMSAYAARRLANETPAQRQARLLRMSAYAAKRQAS
H. sapiens EPLEVRLQRLERERTAKKSRRDNETPEEREVRRMRDREAKRLQRMQETDEQRARRLQRDREAMRLKRANETPEKRQARLIREREAKRLKRRL
D. rerio EPVEVRMQRLERERTAKKTKRACETPEEREMRRLRDREAKRMQRMQESDEQRSRRLQRDREAMRLKRANETPEKRQARLIREREAKRLKRRL
Drosophila-1 ESIEARERRLERNAARMRDKRAKESEAEYRVRLAKNAEANRVRRQNETEVQRTLRLMKNAARQRLRRASETVEERKKRLAKAAERMRIARAS
Drosophila-2 ESDIERAMRLVRNAARQRLRRAMESPDQRAKRLAKLAERMRMYRASETNDQRKLRLAEMAARARQRIANESPEERKERLRKLNDYAKKVREK
C. elegans ETSEETHTRLAKAADYKRKRCSEETEEEYRERLKRQAERKKMERMQESEDQANNRRIKAAEHRRQYLQRESFEKAQERRNKDAERRRKSLLA
P. patens ESAEQASERRRVEAEKRARLRAGESEEERLQRLRHHAAHQAKKRAEESIEERTERLRKHAAHQAKKRAEESLEERTERLRKHAALQARKRAG

Figure 5. Multiple-sequence alignment of *Bm*STPR against its homologs with the programs *Cobalt* (46) and *Espript* (47). The secondary structural elements of *Bm*STPR are displayed at the top. The three conserved residues such as Glu1, Arg9 and Thr/Ser2 in each repeat are labelled with red stars. The STPR domains are from the following sequences (NCBI accession numbers in parentheses): *B. mori* FMBP-1 (NP_001036969.1), *H. sapiens* Zinc finger protein 821 isoform 2 (NP_060000.1), *D. rerio* predicted Zinc finger protein 821-like isoform X1 (XP_005169107.1), *Drosophila*-1 CG14440 isoform A (NP_572343.1), *Drosophila*-2 CG14442 isoform A (NP_572342.1), *C. elegans* protein C05D11.13 (NP_498414.1) and *P. patens* predicted protein (XP_001767050.1). All STPRs cover the four repeats from R1 to R4, except *Drosophila*-2 covers repeats R3–R6.

Table 2. DNA parameters

DNA segment	Pitch	Rp	Rise	Twist	x-Disp	Roll	Incl	Groove width (Å)		*D*
	(Å)	(Å)	(Å)	(°)	(Å)	(°)	(°)	Minor	Major	(Å)
B-DNA	34.0	9.4	3.3–3.4	36.0	0.10	0.6	2.4	5.7	11.7	3.43
*Bm*STPR-13 bp	32.3	9.4 ± 0.9	3.23 ± 0.12	36.0 ± 3.9	0.06 ± 1.1	− 2.1 ± 3.7	− 3.1 ± 6.1	5.0 ± 1.2	12.8 ± 1.1	3.58
*Bm*STPR-18 bp	33.7	9.6 ± 1.0	3.32 ± 0.20	35.5 ± 4.9	− 0.34 ± 1.2	− 0.4 ± 3.8	− 0.6 ± 6.1	5.0 ± 0.8	13.4 ± 1.1	3.58
*Bm*STPR-20 bp	33.8	9.7 ± 0.8	3.32 ± 0.12	35.3 ± 4.6	− 0.41 ± 0.9	− 1.1 ± 3.4	− 1.8 ± 5.5	4.2 ± 1.1	13.2 ± 0.7	3.84
glucocorticoid-DNA	36.1	10.1 ± 1.4	3.32 ± 0.28	33.1 ± 8.8	− 1.57 ± 1.2	4.5 ± 3.1	8.1 ± 5.7	7.7 ± 0.2	12.9 ± 1.1	2.05
Zif268-DNA	36.8	10.0 ± 0.9	3.29 ± 0.25	32.2 ± 5.4	− 1.57 ± 1.0	4.5 ± 2.7	8.0 ± 4.9	7.6 ± 0.2	11.7 ± 1.5	1.66

The DNA sequences are: *Bm*STPR-13 bp, 5′-tttacatagattc-3′; *Bm*STPR-18 bp, 5′-catacatacatacataca-3′; *Bm*STPR-20 bp, 5′-agtatttacatagattcatc-3′; glucocorticoid-DNA, 5′-gatgttctg-3′; Zif268-DNA, 5′-gcgtgggcgt-3′. The parameters include the pitch, the radius of the best-fit cylinder through all the phosphates (*R*p), the rise, the twist, the displacement (x-Disp), the roll, the inclination (Incl), the groove width (minor and major) and relative displacement (*D*). *D* is defined as the previous report (40).

base A/T with a G/C that alters the major groove width resulted in a sharp decrease of *Bm*STPR binding affinity of 30–60-folds, as seen from the affinity comparison of three DNA sequences (No.1, No.57 and No.116, Supplementary Table S3). It indicated that the recognition of *Bm*STPR to DNA is a combination of direct and indirect interactions; however, the main contribution is from the indirect readout.

The DNA geometry in the three complex structures

It was reported that DNA bound to helical proteins in the major groove adopts a deformed B-DNA conformation, for example B_{eg}-DNA (where eg stands for enlarged groove) (40). Using the *3DNA* server (http://w3dna.rutgers.edu/) (41), we performed a DNA geometry analysis of our three DNA structures through nine major parameters (Table 2). Upon binding to *Bm*STPR via the major groove, the three DNA sequences share a structure of quite similar parameters to each other. However, compared to the canonical B-DNA (42), the different values in x-displacement, roll angle, inclination degree and groove width indicated that our three DNA structures adopt a deformed B-DNA conformation induced by *Bm*STPR binding (Table 2). Moreover, the three DNAs exhibit a different structure from the previously defined B_{eg}-DNA (40), which also binds to helical proteins via the enlarged major groove. Compared to the two B_{eg}-DNA representatives glucocorticoid-DNA (PDB code: 1R4O) and Zif268-DNA (PDB code: 1ZAA), *Bm*STPR-bound DNAs have a negative x-displacement and a negative inclination degree, indicating a distinct relative position between base pair and helical axis, in addition to an altered relative displacement, which corresponds to the spatial relationship between the base pairs and the phosphate backbone (Table 2). In addition, our three DNA structures display an average value of negative roll angle, different from that for either the canonical B-DNA or B_{eg}-DNA (Table 2). All together, the three *Bm*STPR-bound DNAs adopt a unique deformed B-DNA conformation which is distinct from the previously defined B_{eg}-DNA. Notably, compared to the 11.7-Å major groove width for canonical B-DNA, the three *Bm*STPR-bound DNAs share a rather wider major groove of 12.8, 13.4 and 13.2 Å in average, respectively (Table 2). In fact, the AT-rich sequences usually adopt a narrower minor groove (28,29), in consequence a wider major groove, as the widths of minor and major grooves are usually correlated to each other (43). Moreover, comparison of the key parameters of the *Bm*STPR-bound DNA structures with the free AT-rich DNA structures (Table 2 and Supplementary Table S4) revealed a significant induced fit upon binding to *Bm*STPR. Together, we propose that the high flexibility and intrinsically wider major groove of AT-rich DNAs contribute to the specific recognition towards *Bm*STPR.

STPR-containing proteins are widely spread in animals

Sequence homology search against the NCBI database (http://blast.ncbi.nih.nlm.gov) (44,45) yielded an output of 178 STPR-containing proteins of a sequence-identity higher than 37% with an *E*-value <80. Similar to *Bm*STPR, most STPR domains consist of four repeats. However, there are a few exceptions that possess three repeats or five to seven repeats. Interestingly, all STPR-containing proteins are mainly distributed in animals, except for one case from *Physcomitrella patens* which possesses five repeats. We aligned the STPR domains of proteins from the model organisms including human, *Caenorhabditis elegans*, *Danio rerio* and *Drosophila melanogaster*, in addition to *P. patens*. Each repeat is strictly composed of 23 residues and rich of basic residues (Figure 5), indicating its DNA-binding capacity. Moreover, each repeat harbours three highly conserved residues: Glu1, Arg9 and Thr/Ser2 (Figure 5), which contribute to stabilizing the α-helical conformation of the N-terminal moiety of each repeat. Thus, we propose that the STPR-containing proteins from other organisms might also be able to wrap the favoured DNA along the major groove in a somewhat similar pattern. However, these proteins are usually fused with various domains either at the N- and/or C-terminus, indicating their diverse physiological functions.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

We appreciate the help of the staff at the Shanghai Synchrotron Radiation Facility.

FUNDING

The 973 project from the Ministry of Science and Technology of China [2012CB114601, 2012CB911002]; Chinese National Natural Science Foundation [31572318, 31272362]; Fundamental Research Funds for the Central Universities [WK2070000054]. Funding for open access charge: The 973 project from the Ministry of Science and Technology of China [2012CB114601, 2012CB911002]; Chinese National Natural Science Foundation [31572318, 31272362]; Fundamental Research Funds for the Central Universities [WK2070000054].

Conflict of interest statement. None declared.

REFERENCES

1. Anderson,J.E., Ptashne,M. and Harrison,S.C. (1987) Structure of the repressor-operator complex of bacteriophage 434. *Nature*, **326**, 846–852.
2. Matthews,B.W. (1988) Protein-DNA interaction. No code for recognition. *Nature*, **335**, 294–295.
3. Pabo,C.O., Aggarwal,A.K., Jordan,S.R., Beamer,L.J., Obeysekare,U.R. and Harrison,S.C. (1990) Conserved residues make similar contacts in two repressor-operator complexes. *Science*, **247**, 1210–1213.
4. Pavletich,N.P. and Pabo,C.O. (1993) Crystal structure of a five-finger GLI-DNA complex: new perspectives on zinc fingers. *Science*, **261**, 1701–1707.
5. Rohs,R., Jin,X., West,S.M., Joshi,R., Honig,B. and Mann,R.S. (2010) Origins of specificity in protein-DNA recognition. *Annu. Rev. Biochem.*, **79**, 233–269.
6. Otwinowski,Z., Schevitz,R.W., Zhang,R.G., Lawson,C.L., Joachimiak,A., Marmorstein,R.Q., Luisi,B.F. and Sigler,P.B. (1988) Crystal structure of trp repressor/operator complex at atomic resolution. *Nature*, **335**, 321–329.
7. Seeman,N.C., Rosenberg,J.M. and Rich,A. (1976) Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl. Acad. Sci. U.S.A.*, **73**, 804–808.
8. Harrison,S.C. and Aggarwal,A.K. (1990) DNA recognition by proteins with the helix-turn-helix motif. *Annu. Rev. Biochem.*, **59**, 933–969.
9. Bewley,C.A., Gronenborn,A.M. and Clore,G.M. (1998) Minor groove-binding architectural proteins: structure, function, and DNA recognition. *Annu. Rev. Biophys. Biomol. Struct.*, **27**, 105–131.
10. Nelson,H.C., Finch,J.T., Luisi,B.F. and Klug,A. (1987) The structure of an oligo(dA).oligo(dT) tract and its biological implications. *Nature*, **330**, 221–226.
11. Hizver,J., Rozenberg,H., Frolow,F., Rabinovich,D. and Shakked,Z. (2001) DNA bending by an adenine-thymine tract and its role in gene regulation. *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 8490–8495.
12. Haran,T.E. and Mohanty,U. (2009) The unique structure of A-tracts and intrinsic DNA bending. *Q. Rev. Biophys.*, **42**, 41–81.
13. Kirsanov,D.D., Zanegina,O.N., Aksianov,E.A., Spirin,S.A., Karyagina,A.S. and Alexeevski,A.V. (2013) NPIDB: nucleic acid-protein interaction database. *Nucleic Acids Res.*, **41**, D517–D523.
14. Zanegina,O., Kirsanov,D., Baulin,E., Karyagina,A., Alexeevski,A. and Spirin,S. (2016) An updated version of NPIDB includes new classifications of DNA-protein complexes and their families. *Nucleic Acids Res.*, **44**, D144–D153.
15. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
16. Inoue,S., Tanaka,K., Arisaka,F., Kimura,S., Ohtomo,K. and Mizuno,S. (2000) Silk fibroin of *Bombyx mori* is secreted, assembling a high molecular mass elementary unit consisting of H-chain, L-chain, and P25, with a 6:6:1 molar ratio. *J. Biol. Chem.*, **275**, 40517–40528.
17. Suzuki,Y. and Giza,P.E. (1976) Accentuated expression of silk fibroin genes in vivo and in vitro. *J. Mol. Biol.*, **107**, 183–206.
18. Hui,C.C., Matsuno,K. and Suzuki,Y. (1990) Fibroin gene promoter contains a cluster of homeodomain binding sites that interact with three silk gland factors. *J. Mol. Biol.*, **213**, 651–670.
19. Suzuki,T., Matsuno,K., Takiya,S., Ohno,K., Ueno,K. and Suzuki,Y. (1991) Purification and characterization of an enhancer-binding protein of the fibroin gene. I. Complete purification of fibroin factor 1. *J. Biol. Chem.*, **266**, 16935–16941.
20. Suzuki,T., Takiya,S., Matsuno,K., Ohno,K., Ueno,K. and Suzuki,Y. (1991) Purification and characterization of an enhancer-binding protein of the fibroin gene. II. Functional analyses of fibroin factor 1. *J. Biol. Chem.*, **266**, 16942–16947.
21. Takiya,S., Kokubo,H. and Suzuki,Y. (1997) Transcriptional regulatory elements in the upstream and intron of the fibroin gene bind three specific factors POU-M1, Bm Fkh and FMBP-1. *Biochem. J.*, **321**, 645–653.
22. Suzuki,Y., Tsuda,M., Hirose,S. and Takiya,S. (1986) Transcription signals and factors of the silk genes. *Adv. Biophys.*, **21**, 205–215.
23. Maekawa,H. and Suzuki,Y. (1980) Repeated turn-off and turn-on of fibroin gene transcription during silk gland development of *Bombyx mori*. *Dev. Biol.*, **78**, 394–406.
24. Takiya,S., Ishikawa,T., Ohtsuka,K., Nishita,Y. and Suzuki,Y. (2005) Fibroin-modulator-binding protein-1 (FMBP-1) contains a novel DNA-binding domain, repeats of the score and three amino acid peptide (STP), conserved from *Caenorhabditis elegans* to humans. *Nucleic Acids Res.*, **33**, 786–795.
25. Saito,S., Aizawa,T., Kawaguchi,K., Yamaki,T., Matsumoto,D., Kamiya,M., Kumaki,Y., Mizuguchi,M., Takiya,S., Demura,M. *et al.* (2007) Structural approach to a novel tandem repeat DNA-binding domain, STPR, by CD and NMR. *Biochemistry*, **46**, 1703–1713.
26. Saito,S., Yokoyama,T., Aizawa,T., Kawaguchi,K., Yamaki,T., Matsumoto,D., Kamijima,T., Kamiya,M., Kumaki,Y., Mizuguchi,M. *et al.* (2008) Structural properties of the DNA-bound form of a novel tandem repeat DNA-binding domain, STPR. *Proteins*, **72**, 414–426.
27. Takiya,S., Saito,S., Yokoyama,T., Matsumoto,D., Aizawa,T., Kamiya,M., Demura,M. and Kawano,K. (2009) DNA-binding property of the novel DNA-binding domain STPR in FMBP-1 of the silkworm *Bombyx mori*. *J. Biochem.*, **146**, 103–111.
28. Aymami,J., Nunn,C.M. and Neidle,S. (1999) DNA minor groove recognition of a non-self-complementary AT-rich sequence by a tris-benzimidazole ligand. *Nucleic Acids Res.*, **27**, 2691–2698.
29. Gordon,B.R., Li,Y., Cote,A., Weirauch,M.T., Ding,P., Hughes,T.R., Navarre,W.W., Xia,B. and Liu,J. (2011) Structural basis for recognition of AT-rich DNA by unrelated xenogeneic silencing proteins. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 10690–10695.
30. Otwinowski,Z. and Minor,W. (1997) Processing of X-ray diffraction data collected in oscillation mode. *Method Enzymol.*, **276**, 307–326.
31. Brodersen,D.E., de La Fortelle,E., Vonrhein,C., Bricogne,G., Nyborg,J. and Kjeldgaard,M. (2000) Applications of single-wavelength anomalous dispersion at high and atomic resolution. *Acta Crystallogr. D Biol. Crystallogr.*, **56**, 431–441.
32. Adams,P.D., Grosse-Kunstleve,R.W., Hung,L.W., Ioerger,T.R., McCoy,A.J., Moriarty,N.W., Read,R.J., Sacchettini,J.C., Sauter,N.K. and Terwilliger,T.C. (2002) PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr. D Biol. Crystallogr.*, **58**, 1948–1954.
33. Emsley,P. and Cowtan,K. (2004) Coot: model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.*, **60**, 2126–2132.
34. Murshudov,G.N., Vagin,A.A. and Dodson,E.J. (1997) Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D Biol. Crystallogr.*, **53**, 240–255.
35. Painter,J. and Merritt,E.A. (2006) Optimal description of a protein structure in terms of multiple groups undergoing TLS motion. *Acta Crystallogr. D Biol. Crystallogr.*, **62**, 439–450.
36. Davis,I.W., Leaver-Fay,A., Chen,V.B., Block,J.N., Kapral,G.J., Wang,X., Murray,L.W., Arendall,W.B. III, Snoeyink,J., Richardson,J.S. *et al.* (2007) MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res.*, **35**, W375–W383.
37. Laskowski,R.A., Macarthur,M.W., Moss,D.S. and Thornton,J.M. (1993) Procheck - a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.*, **26**, 283–291.
38. DeLano,W.L. (2002) *The PyMOL Molecular Graphics System*. DeLano Scientific LLC, San Carlos, CA.
39. Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.
40. Nekludova,L. and Pabo,C.O. (1994) Distinctive DNA conformation with enlarged major groove is found in Zn-finger-DNA and other protein-DNA complexes. *Proc. Natl. Acad. Sci. U.S.A.*, **91**, 6948–6952.

41. Zheng,G.H., Lu,X.J. and Olson,W.K. (2009) Web 3DNA-a web server for the analysis, reconstruction, and visualization of three-dimensional nucleic-acid structures. *Nucleic Acids Res.*, **37**, W240–W246.
42. Grasby,J.A., Neidle,S., Blackburn,G.M., Gait,M.J., Loakes,D., Williams,D.M., Egli,M., Flavell,M., Flavell,A. and Pyle,A.M. (2006) DNA and RNA structure. In: Blackburn,GM, Gait,MJ, Loakes,D and Willams,DM (eds). *Nucleic Acids in Chemistry and Biology*. 3rd edn. RSC Publishing, Cambridge, Vol. **3**, pp. 29–30.
43. Boutonnet,N., Hui,X.W. and Zakrzewska,K. (1993) Looking into the grooves of DNA. *Biopolymers*, **33**, 479–490.
44. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J.H., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
45. Johnson,M., Zaretskaya,I., Raytselis,Y., Merezhuk,Y., McGinnis,S. and Madden,T.L. (2008) NCBI BLAST: a better web interface. *Nucleic Acids Res.*, **36**, W5–W9.
46. Papadopoulos,J.S. and Agarwala,R. (2007) COBALT: constraint-based alignment tool for multiple protein sequences. *Bioinformatics*, **23**, 1073–1079.
47. Robert,X. and Gouet,P. (2014) Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res.*, **42**, W320–W324.